Responses to Reviewer 1's comments

The reviewer's comments are given in *black italics* and our responses in blue plain text.

This manuscript provides an extensive intercomparison of diagnostics relevant for polar stratospheric ozone processing in five recent 'full-input' reanalyses, MERRA, MERRA2, CFSR, ERA Interim, and JRA55, as part of the S-RIP intercomparison project. The study is thorough, well thought out and generally clearly presented, and the intercomparison should provide a valuable reference point for studies of polar processing that are based on reanalysis data, as well as a reference point for comparisons of these quantities in future reanalyses. To me the more interesting results are the almost ubiquitous improvement seen in the agreement between reanalyses following the advent of improved satellite observations around 1998-2000, as well as the increased sensitivity to threshold definitions seen in the NH relative to the SH. The results are not earth shattering, but are of value and as such I would recommend publication after some minor revisions.

My main concern is that the paper is very long, and that its impact would be greater if it were significantly shorter. As a potential reference for future studies, there is some value in being rather complete in the intercomparisons, but 21 figures is a lot more than most readers will want to go through. It's not clear to me that Figs. 1-3 are really necessary, nor what is the additional gain from including Figs 18-19 over the content of Figs. 16-17.

We thank the reviewer for their helpful comments. The paper has been extensively revised in response to major comments by the other reviewer, Simon Chabrillat, so there is not a one-to-one correspondence with all of the specific suggestions made by the reviewer. We have, however, tried to keep specific new material as concise as possible and have removed material where it was suggested by either reviewer, as detailed in the specific responses below. This includes removing the original Figures 2, 3, 16, and 17 and the associated discussion.

Two major changes to the paper motivated by Simon Chabrillat's comments are to use a reanalysis ensemble mean (REM) as a reference for the comparisons rather than using MERRA-2 (see our response to Simon for discussion of this), and to remove MERRA from the reanalyses evaluated in this paper. There are numerous reasons for removing the MERRA comparisons, including the following: The choices that were made by GMAO of which products to archive for MERRA have made "fair" comparisons difficult to impossible for many products, including potential vorticity (PV), which is critical for stratospheric vortex and many other studies. While comparing MERRA with MERRA-2 and other reanalyses was critical to evaluating MERRA-2, numerous such studies have now been done; MERRA-2 was intended to supercede MERRA and sufficient evaluation of it has been done now to warrant this. Finally, especially when using the REM as a reference, it is somewhat problematic to include two reanalyses based on nearly the same model in a comparison of just five reanalyses.

Because these two major changes, especially the switch to using the REM, necessitated a nearly complete rewrite of large portions of the text in the results section (though the final results changed very little), several of the reviewers' comments now refer to text that has been replaced, and it is not possible to document every change in detail.

Specific comments
p2 l1 There is a spurious 'data' here.

Fixed.

p2 l32: 'Best' is highly debatable here. They are a good tool, certainly, but they are not appropriate for all tasks.

We have changed this to "among the best".

p7 l6: The role of radiosondes should not be understated here – although it is not considered here, JRA55C, which assimilates only 'conventional' (non-satellite-based) observations does a remarkably good job of capturing much of the details of NH stratospheric variability.

We have added a sentence noting the importance of radiosonde inputs in the lower stratosphere, but also noting the caveat that the sonde data are sparse in the NH polar regions and very sparse in the SH polar regions.

p10 l4-19: The choice of a 5 day geometric mean here needs to be justified here. The key question is the decorrelation timescale of fluctuations in the differences between reanalyses. these could arise from a variety of processes with rather different timescales so it's not at all obvious to me what timescale is appropriate, but given that fluctuations in the physical quantities themselves (temperatures, PV) can have decorrelation timescales of far greater than 5 days this choice could be rendering the derived CIs rather meaningless. This can be checked directly by looking at the autocorrelation functions of some sample quantities.

There is also a question of just what it means for two reanalyses to be 'statistically' indistinguishable. There is an important distinction to be drawn as to whether a difference seen between two temporal averages is indicative of a systematic, steady bias between the two systems as opposed to a result of the residual over temporal fluctuations. But given that these systems are meant to capture the same atmospheric fluctuations, time-dependent differences between reanalyses are still meaningful and potentially quite relevant to know about. Just because this measure indicates that the fluctuations are of larger amplitude than the mean bias (in some statistically meaningful sense) doesn't mean the reanalysis products are indistinguishable.

We have added justification for our choice of the expected block length for the stationary resampling procedure.

Since we moved to using a reanalysis ensemble mean (REM) based on Simon Chabrillat's review, we examined the autocorrelation functions (ACFs) for the differences of the reanalyses from the REM. What we found is that the decorrelation timescales can vary and depend highly on the reanalysis, the diagnostic, the year, and the vertical level; in some cases the decorrelation timescales reach zero in a few days, while in other cases they remain well above zero beyond 10 days. As examples of this, we have attached two figures of the type we used to evaluate these timescales at the end of our responses to reviewer 1. They are large and unwieldy figures, but they show (1) the ACF of the raw diagnostics for the REM (top panel) and the comparison reanalysis (second panel; in these cases MERRA-2), (2) the ACF of the comparison reanalysis minus REM (third panel), and (3) 18 ACFs of 18 different stationary resampled (with expected block length of 5 days) difference time series. The two examples we show here are for SH maximum PV gradients for the same level (490 K) separated by just one year. You can see that for 2015, the autocorrelation of the difference time series (3rd panel) stays fairly large out well beyond 10 days; in contrast, the ACF of the difference time series for 2014 drops much faster. You can also see that even though the decorrelation time scale is quite long for 2015 and the average block length of the resampled time series is 5 days, there are still a handful of resampled cases that also have relatively long decorrelation timescales (see e.g., n = 3, 9, 12, 16, 17, and 18) -- and there are also many resampled time series for 2014 that match the much shorter decorrelation time-scale pretty well too. This is one of the benefits of using the stationary resampling procedure rather than block resampling; using random block sizes can help to create artificial time series that better match the autocorrelation "structure" of the original time series.

After making and examining these sorts of plots, we repeated our bootstrapping procedure and tested using different expected block lengths between and including 5 and 15 days. What we found is that in all cases, the results we obtained were virtually identical. Ultimately, for the results now shown in the manuscript, we increased the expected block length to 10 days since it seemed to be the most "happy medium" among the many ACFs we examined; we also doubled the number of resamples for our bootstrap distributions to 2x10^5.

Regarding your second point, we agree that our results from the bootstrapping analysis should not be used to judge the (in)distinguishability of the reanalyses, but should be limited to the "classical" interpretation of statistical hypothesis testing. The presence of an "x" on our pixel plots (null hypothesis can't be rejected) does not mean that the time series of the certain diagnostic, year, and level are indistinguishable, just that we cannot reject that the winter means are equal. Conversely, the absence of an "x" on our pixel plots (null hypothesis rejected) does not mean that there are overwhelming or large biases, just that the winter means are unlikely to be equal. The significance testing here primarily supplements the winter mean differences and standard deviations -- for example, there are many cases of the diagnostic mean differences being very small but "significant" (no "x") alongside standard deviations that are very small,

which just says that although such differences are generally small, they are persistent enough during the season such that many resamples of the time series shared that persistent (but small) difference. There are also some cases where the diagnostic mean differences are noticeably nonzero but “insignificant” (“x” is present) alongside larger standard deviations, which indicates that the variability is large enough such that many resamples do not share the structures that give rise to the real mean difference.

We have modified and double checked our text to ensure we have not included any misleading language regarding the interpretation of the statistics.

p10 l22-24: Are these averages and standard deviations taken over time (from the 12Z snapshots) within the year? Or are they taken over spatial degrees of freedom? Is the data synthetic? If not, what is actually shown?

A more general thought on this section - while I appreciate the effort to make the plots clear I wonder if it would be more efficient to simply explain this plot in the first case rather than present an example; the paper is quite long and omitting Figures 2 and 3 would go some ways towards shortening it without omitting relevant details.

The data here were synthetic and meant to represent averages and standard deviations taken over time as in the other results we show, but we have taken your advice to shorten the paper and have ultimately taken out (what were formerly) Figures 2 and 3.

Fig. 4: What is the relevance of the black lines 70 hPa and 30 hPa?

These are the selected levels for which separate line plots are shown. This is now clarified in the caption.

Fig. 5: Four digits of precision are not needed on the pressure axis labels

The labels are now limited to a single digit after the decimal point in all the figures.

p13 l3: Earlier in the text A_PSC has been used - this to my mind is more standard than A_NAT. Was the switch intentional?

We use the subscripts _NAT and _ice to convey the particular type of PSC threshold we are looking at. A note to this effect has been added in discussion of PSC thresholds in the methods section.

p15 l34: Up to 600K or so there is a significant improvement in the agreement between MERRA and MERRA 2 (in means and standard deviations) after 2000 - it’s just in the upper stratosphere (particularly 660 and 700K) that the disagreement becomes if anything larger.

The MERRA comparisons have been removed from the paper for the reasons stated in our response to Simon Chabrillat, so this text has been removed.

p16 l2: Is this a result of a more or less constant PV offset across the polar regions or differences in the locations of the maximum gradient?

This text has been revised to reflect the individual calculations of the vortex edge location for each reanalysis. The results now suggest that this is related to differences in the locations of the maximum PV gradients, which is noted in the revised text.

p16 l23: 'Total days' is a strange unit here since it's regularly far in excess of the total number of days in a year. The appropriate unit should be pressure-level days, I suppose.

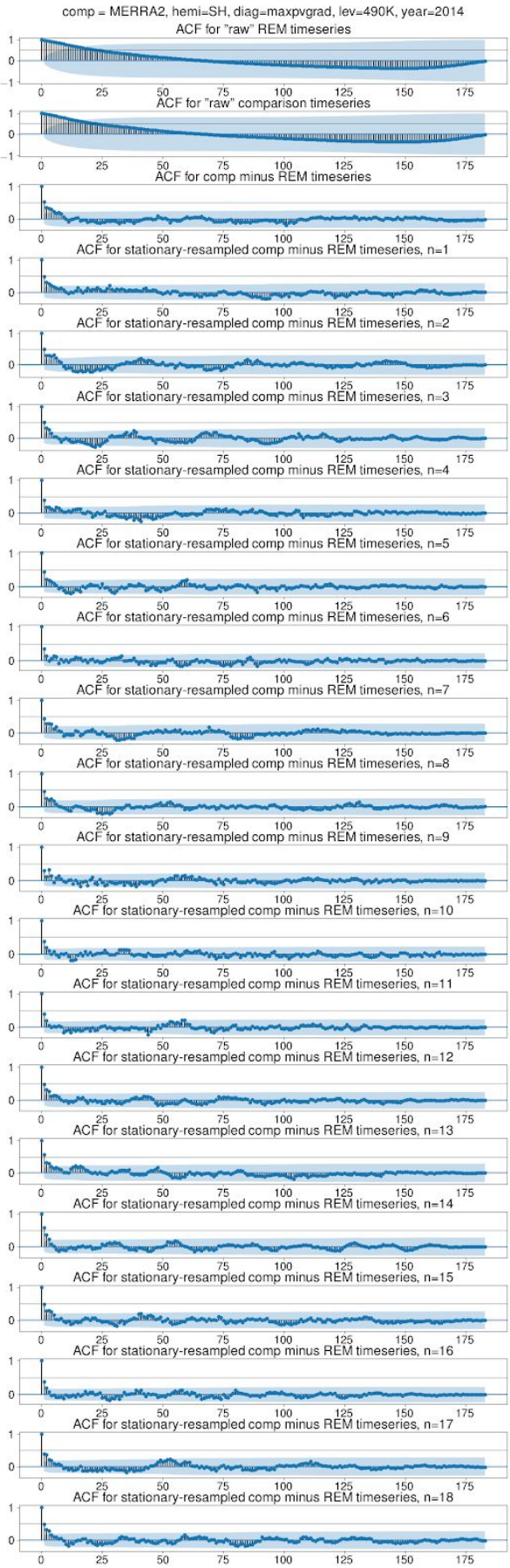
Because the V_PSC / V_vort figures provide much of the same information, and to shorten the paper, we have followed your suggestion to delete the plots showing days integrated over the levels, so these figures have been removed.

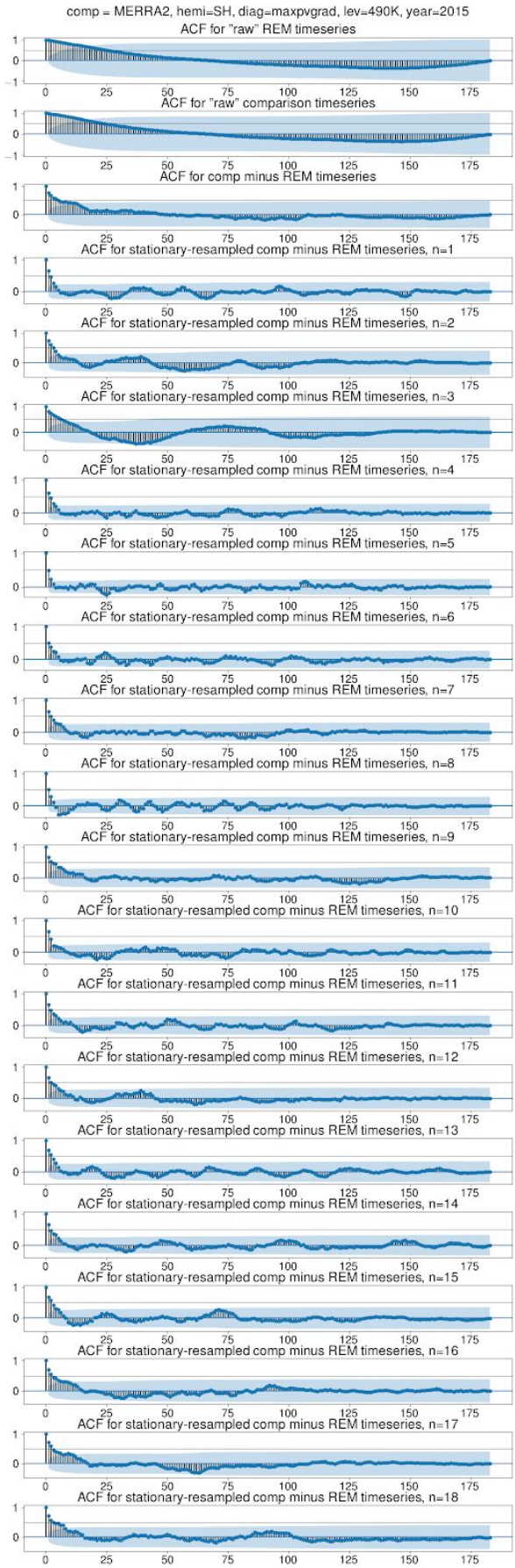
p18 l26: I can't find an explicit definition of A_vort, though there are some relevant details in section 2.2.2

Since we do not use "A_vort" elsewhere in the paper, we now simply refer to it as "vortex area". We have also made the definition of vortex area more explicit. However, please note that the paragraphs discussing the methods behind the derived diagnostics have been moved to a new subsubsection of section 2 (in response to a comment by Simon Chabrillat).

p22 l34: Given the statement two lines earlier about the similar timing of changes in the observations being assimilated by different reanalyses, the consistency of trends across multiple reanalyses should not be seen as any kind of definitive indication of the reliability of trends.

Agreement across reanalyses would be a ***necessary*** condition to believe trends derived from them to be reliable. We agree that it is certainly not a ***sufficient*** condition. We have reworded the sentence in question to make this more explicit.


comp = MERRA2, hemi=SH, diag=maxpvgrad, lev=490K, year=2014
ACF for "raw" REM timeseries
ACF for "raw" comparison timeseries
ACF for comp minus REM timeseries
ACF for stationary-resampled comp minus REM timeseries, n=1
ACF for stationary-resampled comp minus REM timeseries, n=2
ACF for stationary-resampled comp minus REM timeseries, n=3
ACF for stationary-resampled comp minus REM timeseries, n=4
ACF for stationary-resampled comp minus REM timeseries, n=5
ACF for stationary-resampled comp minus REM timeseries, n=6
ACF for stationary-resampled comp minus REM timeseries, n=7
ACF for stationary-resampled comp minus REM timeseries, n=8
ACF for stationary-resampled comp minus REM timeseries, n=9
ACF for stationary-resampled comp minus REM timeseries, n=10
ACF for stationary-resampled comp minus REM timeseries, n=11
ACF for stationary-resampled comp minus REM timeseries, n=12
ACF for stationary-resampled comp minus REM timeseries, n=13
ACF for stationary-resampled comp minus REM timeseries, n=14
ACF for stationary-resampled comp minus REM timeseries, n=15
ACF for stationary-resampled comp minus REM timeseries, n=16
ACF for stationary-resampled comp minus REM timeseries, n=17
ACF for stationary-resampled comp minus REM timeseries, n=18


comp = MERRA2, hemi=SH, diag=maxpvgrad, lev=490K, year=2015
ACF for "raw" REM timeseries
ACF for "raw" comparison timeseries
ACF for comp minus REM timeseries
ACF for stationary-resampled comp minus REM timeseries, n=1
ACF for stationary-resampled comp minus REM timeseries, n=2
ACF for stationary-resampled comp minus REM timeseries, n=3
ACF for stationary-resampled comp minus REM timeseries, n=4
ACF for stationary-resampled comp minus REM timeseries, n=5
ACF for stationary-resampled comp minus REM timeseries, n=6
ACF for stationary-resampled comp minus REM timeseries, n=7
ACF for stationary-resampled comp minus REM timeseries, n=8
ACF for stationary-resampled comp minus REM timeseries, n=9
ACF for stationary-resampled comp minus REM timeseries, n=10
ACF for stationary-resampled comp minus REM timeseries, n=11
ACF for stationary-resampled comp minus REM timeseries, n=12
ACF for stationary-resampled comp minus REM timeseries, n=13
ACF for stationary-resampled comp minus REM timeseries, n=14
ACF for stationary-resampled comp minus REM timeseries, n=15
ACF for stationary-resampled comp minus REM timeseries, n=16
ACF for stationary-resampled comp minus REM timeseries, n=17
ACF for stationary-resampled comp minus REM timeseries, n=18